

Crowdsourcing Ground Truth for Medical Relation Extraction

ANCA DUMITRACHE, Vrije Universiteit Amsterdam, CAS IBM Netherlands
 LORA AROYO, Vrije Universiteit Amsterdam
 CHRIS WELTY, Google Research

Cognitive computing systems require human labeled data for evaluation, and often for training. The standard practice used in gathering this data minimizes disagreement between annotators, and we have found this results in data that fails to account for the ambiguity inherent in language. We have proposed the CrowdTruth method for collecting ground truth through crowdsourcing, that reconsiders the role of people in machine learning based on the observation that disagreement between annotators provides a useful signal for phenomena such as ambiguity in the text. We report on using this method to build an annotated data set for medical relation extraction for the *cause* and *treat* relations, and how this data performed in a supervised training experiment. We demonstrate that by modeling ambiguity, labeled data gathered from crowd workers can (1) reach the level of quality of domain experts for this task while reducing the cost, and (2) provide better training data at scale than distant supervision. We further propose and validate new weighted measures for precision, recall, and F-measure, that account for ambiguity in both human and machine performance on this task.

CCS Concepts: • **Information systems** → **Crowdsourcing**; • **Computing methodologies** → **Language resources**; *Natural language processing*;

General Terms: Human Factors, Experimentation, Performance

Additional Key Words and Phrases: Ground truth, relation extraction, clinical natural language processing, natural language ambiguity, inter-annotator disagreement, CrowdTruth, Crowd Truth

ACM Reference Format:

Anca Dumitrache, Lora Aroyo and Chris Welty, 2016. Crowdsourcing Ground Truth for Medical Relation Extraction. *ACM Trans. Interact. Intell. Syst.* V, N, Article A (January YYYY), 18 pages.
 DOI: <http://dx.doi.org/10.1145/0000000.0000000>

1. INTRODUCTION

Note to reviewers: *The parts of this paper describing the background, motivation, experimental methodology, and some of the experimental results (comparing the crowd-sourced cause relation with expert annotation over 902 sentences), have been previously published in workshop papers [Dumitrache et al. 2015a; 2015b]. This submission synthesizes those two papers while adding significant new contributions. As a result, we have re-used some of the text, tables and figures from the workshop papers for the sake of keeping this paper more complete and self-contained, and describe several significant new contributions that have never been published before:*

- *the dataset in the experiment has been extended to 3,984 sentences, from 902 (Fig.9 & Fig.10 & Tab.IV), to demonstrate the scalability of our approach and the stability of the classifier performance points we used for comparison*

Author's addresses: A. Dumitrache and L. Aroyo, Business Web & Media Department, Vrije Universiteit Amsterdam; C. Welty, Google Research New York.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© YYYY ACM. 2160-6455/YYYY/01-ARTA \$15.00

DOI: <http://dx.doi.org/10.1145/0000000.0000000>

- *all the experiments were replicated for a second medical relation, treat, in addition to the cause relation, with similar results (Figs.4 & 6, Tab.III), adding further evidence to validate the approach and contributing a significant new resource to the community.*
- *each of these additions adds new analysis and discussion.*

We believe these additional contributions serve to convincingly validate our previous experimental results and bring the work to journal quality.

Many methods for Natural Language Processing (NLP) rely on *gold standard* annotations, or *ground truth*, for the purpose of training, testing and evaluation. Understanding *the role of people in machine learning* is crucial in this context, as human annotation is considered the most reliable method for collecting ground truth. In clinical NLP and other difficult domains, researchers assume that expert knowledge of the field is required from annotators. This means that, aside from the monetary costs of hiring humans to label data, simply finding suitable annotators bears a big time cost. The lack of annotated datasets for training and benchmarking is considered one of the big challenges of clinical NLP [Chapman et al. 2011].

Furthermore, the *standard data labeling practice used in supervised machine learning* often presents flaws. Data labeling is performed by humans, by reading text and following a set of guidelines to ensure a uniform understanding of the annotation task. It is assumed that the gold standard represents a universal and reliable model for language. However, [Schaeckermann et al. 2016] and [Bayerl and Paul 2011] criticize this approach by investigating the role of inter-annotator disagreement as a possible indicator of ambiguity inherent in text. Previous experiments we performed in medical relation extraction [Aroyo and Welty 2013a] support this view by identifying two issues with the standard data labeling practice:

- (1) disagreement between annotators is usually eliminated through overly prescriptive annotation guidelines, thus creating artificial data that is neither general nor reflects the ambiguity inherent in natural language,
- (2) the process of acquiring ground truth by working exclusively with domain experts is costly and non-scalable, both in terms of time and money.

Ambiguity in text also impacts automated processes for extracting ground truth. Specifically, in the case of relation extraction from natural text, distant supervision [Mintz et al. 2009; Welty et al. 2010] is a well-established semi-supervised method that uses pairs of entities known to form a relation (e.g. from a knowledge base) to extract relations from text. However, this approach is also prone to generating low quality data, as not every mention of an entity pair in a sentence means a relation is also present. The problems are further compounded when dealing with ambiguous entities, or incompleteness in the knowledge base.

To address these issues, we propose the *CrowdTruth* method for crowdsourcing training data for machine learning. We present an alternative approach for guiding supervised machine learning systems beyond the standard data labeling practice of a universal ground truth, by instead harnessing disagreement in crowd annotations to model the ambiguity inherent in text. We claim that, even for complex annotation tasks such as relation extraction, lack of domain expertise of the crowd is compensated by collecting a large enough set of annotations.

Previously, we studied medical relation extraction in a relatively small set of 90 sentences [Aroyo and Welty 2013b], comparing the results from the crowd with that of two expert medical annotators. We found that disagreement within the crowd is consistent with expert inter-annotator disagreement. Furthermore, sentences that regis-

tered high disagreement tended to be vague or ambiguous when manually evaluated. In this paper, we build on these results by training a classifier for medical relation extraction with CrowdTruth data, and evaluating its performance. The goal is to show that harnessing inter-annotator disagreement results in improved performance for relation extraction classifiers. Our contributions are the following:

- (1) a comparison between using annotations from crowd and from medical experts to train a relation extraction classifier, showing that, with the processing of ambiguity, *classifiers trained on crowd annotations perform the same as to those trained on experts*;
- (2) a similar comparison between crowd annotations and distant supervision, showing that *classifiers trained on crowd annotations perform better than those trained on distant supervision*;
- (3) a *dataset of 3,984 English sentences for medical relation extraction*, centering on the *cause* and *treat* relations, that have been processed with disagreement analysis to capture ambiguity, openly available at:
<https://github.com/CrowdTruth/Medical-Relation-Extraction>.

2. RELATED WORK

2.1. Medical crowdsourcing

There exists some research using crowdsourcing to collect semantic data for the medical domain. [Mortensen et al. 2013] use crowdsourcing to verify relation hierarchies in biomedical ontologies. On 14 relations from the SNOMED CT CORE Problem List Subset, the authors report the crowd's accuracy at 85% for identifying whether the relations were correct or not. In the field of Biomedical NLP, [Burger et al. 2012] used crowdsourcing to extract the gene-mutation relations in Medical Literature Analysis and Retrieval System Online (MEDLINE) abstracts. Focusing on a very specific gene-mutation domain, the authors report a weighted accuracy of 82% over a corpus of 250 MEDLINE abstracts. Finally, [Li et al. 2015] performed a study exposing ambiguities in a gold standard for drug-disease relations with crowdsourcing. They found that, over a corpus of 60 sentences, levels of crowd agreement varied in a similar manner to the levels of agreement among the original expert annotators. All of these approaches present preliminary results from experiments performed with small datasets.

To our knowledge, the most extensive study of medical crowdsourcing was performed by [Zhai et al. 2013], who describe a method for crowdsourcing a ground truth for medical named entity recognition and entity linking. In a dataset of over 1,000 clinical trials, the authors show no statistically significant difference between the crowd and expert-generated gold standard for the task of extracting medications and their attributes. We extend these results by applying crowdsourcing to the more complex task of medical relation extraction, that *prima facie* seems to require more domain expertise than named entity recognition. Furthermore, we test the viability of the crowdsourced ground truth by training a classifier for relation extraction.

2.2. Crowdsourcing ground truth

Crowdsourcing ground truth has shown promising results in a variety of other domains. [Snow et al. 2008] have shown that aggregating the answers of an increasing number of unskilled crowd workers with majority vote can lead to high quality NLP training data. [Hovy et al. 2014] compared the crowd versus experts for the task of part-of-speech tagging. The authors also show that models trained based on crowd-sourced annotation can perform just as well as expert-trained models. [Kondreddi et al. 2014] studied crowdsourcing for relation extraction in the general domain, comparing its efficiency to that of fully automated information extraction approaches. Their

results showed the crowd was especially suited to identifying subtle formulations of relations that do not appear frequently enough to be picked up by statistical methods.

Other research for crowdsourcing ground truth includes: entity clustering and disambiguation [Lee et al. 2013], Twitter entity extraction [Finin et al. 2010], multilingual entity extraction and paraphrasing [Chen and Dolan 2011], and taxonomy creation [Chilton et al. 2013]. However, all of these approaches rely on the assumption that one black-and-white gold standard must exist for every task. Disagreement between annotators is discarded by picking one answer that reflects some consensus, usually through using majority vote. The number of annotators per task is also kept low, between two and five workers, in the interest of reducing cost and eliminating disagreement. [Whitehill et al. 2009] and [Welinder et al. 2010] have used a latent variable model for task difficulty, as well as latent variables to measure the skill of each annotator, to optimize crowdsourcing for image labels. The novelty in our approach is to consider language ambiguity, and consequently inter-annotator disagreement, as an inherent feature of the language. Language ambiguity can be related to, but is not necessarily a direct cause of task difficulty. The metrics we employ for determining the quality of crowd answers are specifically tailored to measure ambiguity by quantifying disagreement between annotators.

2.3. Disagreement and ambiguity in crowdsourcing

In addition to our own work [Aroyo and Welty 2013a], the role of ambiguity when building a gold standard has previously been discussed by [Lau et al. 2014]. The authors propose a method for crowdsourcing ambiguity in the grammatical correctness of text by giving workers the possibility to pick various degrees of correctness. However, inter-annotator disagreement is not discussed as a factor in measuring this ambiguity. After empirically studying part-of-speech datasets, [Plank et al. 2014] found that inter-annotator disagreement is consistent across domains, even across languages. Furthermore, most disagreement is indicative of debatable cases in linguistic theory, rather than faulty annotation. We believe these findings manifest even more strongly for NLP tasks involving semantic ambiguity, such as relation extraction.

In assessing the Ontology Alignment Evaluation Initiative (OAEI) benchmark, [Cheatham and Hitzler 2014] found that disagreement between annotators (both crowd and expert) is an indicator for inherent uncertainty in the domain knowledge, and that current benchmarks in ontology alignment and evaluation are not designed to model this uncertainty. This position is shared by [Schaekermann et al. 2016], who propose a framework for dealing with uncertainty in ground truth that acknowledges the notion of ambiguity, and uses disagreement in crowdsourcing for modeling this ambiguity. To our knowledge, our work presents the first experimental results of using disagreement-aware crowdsourcing for training a machine learning system.

3. EXPERIMENTAL SETUP

The goal of our experiments is to assess the quality of our disagreement-aware crowdsourced data in training a medical relation extraction model. We use a binary classifier [Wang and Fan 2014] that takes as input a set of sentences and two terms from the sentence, and returns a score reflecting the likelihood that a specific relation is expressed in the sentence between the terms. This manifold learning classifier was one of the first to accept weighted scores for each training instance, although it still requires a discrete positive or negative label. This property seemed to make it suitable for our experiments, as we expected the ambiguity of a sentence to impact its suitability as a training instance (in other words, we decreased the weight of training instances that exhibited ambiguity). We investigate the performance of the classifier over two medi-

cal relations: *cause* (between symptoms and disorders) and *treat* (between drugs and disorders).

Relation	Corresponding UMLS relation(s)	Definition	Example
<i>treat</i>	may treat	therapeutic use of a drug	penicillin treats infection
<i>cause</i>	cause of; has causative agent	the underlying reason for a symptom or a disease	fever induces dizziness
<i>prevent</i>	may prevent	preventative use of a drug	vitamin C prevents influenza
<i>diagnoses</i>	may diagnose	diagnostic use of an ingredient, test or a drug	RINNE test is used to diagnose hearing loss
<i>location</i>	disease has primary anatomic site; has finding site	body part in which disease or disorder is observed	leukemia is found in the circulatory system
<i>symptom</i>	disease has finding; disease may have finding	deviation from normal function indicating the presence of disease or abnormality	pain is a symptom of a broken arm
<i>manifestation</i>	has manifestation	links disorders to the observations that are closely associated with them	abdominal distention is a manifestation of liver failure
<i>contraindicate</i>	contraindicated drug	a condition for which a drug or treatment should not be used	patients with obesity should avoid using danazol
<i>side effect</i>	side effect	a secondary condition or symptom that results from a drug	use of antidepressants causes dryness in the eyes
<i>associated with</i>	associated with	signs, symptoms or findings that often appear together	patients who smoke often have yellow teeth
<i>is a</i>	is a	a relation that indicates that one of the terms is more specific variation of the other	migraine is a kind of headache
<i>part of</i>	part of	an anatomical or structural sub-component	the left ventricle is part of the heart

Table I: Set of medical relations.

The quality of the crowd data in training the classifier is evaluated in two parts: first by comparing it to the performance of an expert-trained classifier, and second with a classifier trained on distant supervision data. The training is done separately for each relation, over the same set of sentences, with different relation existence labels for crowd, expert and baseline.

3.1. Data selection

The dataset used in our experiments contains 3,984 medical sentences extracted from PubMed article abstracts. The sentences were sampled from the set collected by [Wang and Fan 2014] for training the relation extraction model that we are re-using. Wang & Fan collected the sentences with *distant supervision* [Mintz et al. 2009; Welty et al. 2010], a method that picks positive sentences from a corpus based on whether known arguments of the seed relation appear together in the sentence (e.g. the *treat* relation occurs between terms *antibiotics* and *typhus*, so find all sentences containing both and repeat this for all pairs of arguments that hold). The MetaMap parser [Aronson 2001] was used to recognize medical terms in the corpus, and the UMLS vocabulary [Bodenreider 2004] was used for mapping terms to categories, and relations to term types. The intuition of distant supervision is that since we know the terms are related, and they are in the same sentence, it is more likely that the sentence expresses a relation between them (than just any random sentence).

We started with a set of 12 relations important for clinical decision making, used also by Wang & Fan. Each of these relations corresponds to a set of UMLS relations (Tab.I), as UMLS relations are sometimes overlapping in meaning (e.g. *cause of* and *has causative agent* both map to *cause*). The UMLS relations were used as a seed in distant supervision. We focused our efforts on the relations *cause* and *treat*. These two relations were used as a seed for distant supervision in two thirds of the sentences of our dataset (1,043 sentences for *treat*, 1,828 for *cause*). The final third of the sentences were collected using the other 10 relations as seeds, in order to make the data more heterogeneous.

To perform a comparison with expert-annotated data, we randomly sampled a set of 902 sentences from the distant supervision dataset. This set restriction was done not just due to the cost of the experts, but primarily because of their limited time and availability. To collect this data, we employed medical students, in their third year at American universities, that had just taken United States Medical Licensing Examination (USMLE) and were waiting for their results. Each sentence was annotated by exactly one person. The annotation task consisted of deciding whether or not the UMLS seed relation discovered by distant supervision is present in the sentence for the two selected terms.

The crowdsourced annotation setup is based on our previous medical relation extraction work [Aroyo and Welty 2014]. For every sentence, the crowd was asked to decide which relations (from Tab.I) hold between the two extracted terms. The task was multiple choice, workers being able to choose more than one relation at the same time. There were also options available for cases when the medical relation was other than the ones we provided (*other*), and for when there was no relation between the terms (*none*). The crowdsourcing was run on the CrowdFlower¹ platform, with 15 workers per sentence. Compared to a single expert judgment per sentence, the total cost of the crowd amounted to 2/3 of the sum paid for the experts.

All of the data that we have used, together with the templates for the crowdsourcing tasks, and the crowdsourcing implementation details are available online at: <https://github.com/CrowdTruth/Medical-Relation-Extraction>.

¹<https://crowdfower.com/>

3.2. CrowdTruth metrics

The crowd output was processed with the use of CrowdTruth metrics – a set of general-purpose crowdsourcing metrics [Inel et al. 2014], that have been successfully used to model ambiguity in annotations for relation extraction, event extraction, sounds, images, and videos [Aroyo and Welty 2014]. These metrics model ambiguity in semantic interpretation based on the triangle of reference [Ogden and Richards 1923], with the vertices being the input sentence, the worker, and the seed relation. Ambiguity and disagreement at any of the vertices (e.g. a sentence with unclear meaning, a poor quality worker, or an unclear relation) will propagate in the system, influencing the other components. For example, if a sentence is unclear, we expect workers will be more likely to disagree with each other; if a worker is not doing a good job, we expect that worker to consistently disagree with other workers across all the sentences they worked on; and if a particular target relation is unclear, we expect workers to disagree on the application of that relation across all the sentences. By using multiple workers per sentence and requiring each worker to annotate multiple sentences, the aggregate data helps us isolate these individual signals and how they interact. Thus a high quality worker who annotates a low clarity sentence will be recognized as high quality. In our workflow, these metrics are used both to eliminate spammers, as detailed by [Aroyo and Welty 2014], and to determine the clarity of the sentences and relations. The main concepts are:

- *annotation vector*: used to model the annotations of one worker for one sentence. For each worker i submitting their solution to a task on a sentence s , the vector $W_{s,i}$ records their answers. If the worker selects a relation, its corresponding component would be marked with ‘1’, and ‘0’ otherwise. The vector has 14 components, one for each relation, as well as *none* and *other*. Multiple choices (e.g. picking multiple relations for the same sentence) are modeled by marking all corresponding vector components with ‘1’.
- *sentence vector*: the main component for modeling disagreement. For every sentence s , it is computed by adding the annotation vectors for all workers on the given task: $V_s = \sum_i W_{s,i}$. One such vector was calculated for every sentence.
- *sentence-relation score*: measures the ambiguity of a specific relation in a sentence with the use of cosine similarity. The higher the score, the more clearly the relation is expressed in the sentence. The sentence-relation score is computed as the cosine similarity between the sentence vector and the unit vector for the relation: $srs(s, r) = \cos(V_s, \hat{r})$, where the unit vector \hat{r} refers to a vector where the component corresponding to relation r is equal to ‘1’, and all other components are equal to ‘0’. The reasoning is that the unit vector \hat{r} corresponds to the clearest representation of a relation in a sentence – i.e. when all workers agree that relation r exists between the seed terms, and all other relations do not exist. As a cosine similarity, these scores are in the $[0, 1]$ interval. Tab.II shows the transformation of sentence vectors to the sentence-relation scores and then to the training scores using the threshold below.
- *sentence-relation score threshold*: a fixed value in the interval $[0, 1]$ used to differentiate between a negative and a positive label for a relation in a sentence. Given a value t for the threshold, all sentences with a sentence-relation score less than t get a negative label, and the ones with a score greater or equal to t are positive. The results section compares the performance of the crowd at different threshold values. This threshold was necessary because our classifier required either a positive or negative label for each training example.

Table II: Given two sentences, *Sent.1* and *Sent.2*, with term pairs in bold font, the table shows the transformation of the sentence vectors to sentence – relation scores, and then to *crowd* model train scores. The sentence-relation threshold for the train score is set at 0.5 for these examples.

Sent.1: **Renal osteodystrophy** is a general complication of chronic renal failure and **end stage renal disease**.

Sent.2: If **TB** is a concern, a **PPD** is performed.

Relation	sentence vector		sentence – relation score		crowd model train score	
	<i>Sent.1</i>	<i>Sent.2</i>	<i>Sent.1</i>	<i>Sent.2</i>	<i>Sent.1</i>	<i>Sent.2</i>
<i>treat</i>	0	3	0	0.36	-1	-0.64
<i>prevent</i>	0	1	0	0.12	-1	-0.88
<i>diagnose</i>	1	7	0.09	0.84	-0.91	0.84
<i>cause</i>	10	0	0.96	0	0.96	-1
<i>location</i>	1	0	0.09	0	-0.91	-1
<i>symptom</i>	2	0	0.19	0	-0.81	-1
<i>manifestation</i>	0	0	0	0	-1	-1
<i>contraindicate</i>	0	0	0	0	-1	-1
<i>associated with</i>	1	3	0.09	0.36	-0.91	-0.64
<i>side effect</i>	0	0	0	0	-1	-1
<i>is a</i>	0	0	0	0	-1	-1
<i>part of</i>	0	0	0	0	-1	-1
<i>other</i>	0	1	0	0.12	-1	-0.88
<i>none</i>	0	0	0	0	-1	-1

3.3. Training the model

The sentences together with the relation annotations were then used to train a manifold model for relation extraction [Wang and Fan 2014]. This model was developed for the medical domain, and tested for the relation set that we employ. It is trained per individual relation, by feeding it both *positive* and *negative* data. It offers support for both discrete labels, and real values for weighting the confidence of the training data entries, with positive values in $(0, 1]$, and negative values in $[-1, 0)$. Using this system, we train several models using five-fold cross validation, in order to assess the performance of the crowd dataset. The training was done separately for the *treat* and *cause* relations. For each relation, we constructed four datasets, with the same sentences and term pairs, but with different labels for whether or not the relation is present in the sentence:

- (1) *baseline*: The distant supervision data is used to provide discrete (positive or negative) labels on each sentence - i.e. if a sentence contains two terms known (in UMLS) to be related by *treats*, the sentence is considered positive. Distant supervision does not extract negative examples, so in order to generate a negative set for one relation, we use positive examples for the other (non-overlapping) relations shown in Tab. I. This dataset constitutes the baseline against which all other datasets are tested.
- (2) *expert*: Discrete labels based on an expert's judgment as to whether the *baseline* label is correct. The experts do not generate judgments for all combinations of sentences and relations – for each sentence, the annotator decides on the seed relation extracted with distant supervision. Similarly to the baseline data, we reuse positive examples from the other relations to increase the number of negative examples.

- (3) *single*: Discrete labels for every sentence are taken from one randomly selected crowd worker who annotated the sentence. This data simulates the traditional single annotator setting common in annotation environments.
- (4) *crowd*: Weighted labels for every sentence are based on the CrowdTruth *sentence-relation score*. The classifier expects positive scores for positive examples, and negative scores for negative, so the sentence-relation scores must be re-scaled in the $[-1, 0]$ interval for negative labels. An example of how the scores were processed is given in Tab.II.

For each relation, two experiments were run. First, we performed a comparison between the *crowd* and *expert* datasets by training a model using the subset of 902 sentences that also has expert annotations. Next, after we were able to determine the quality of the *crowd* data, we performed a second experiment comparing the performance of the classifier when trained with the *crowd* and *baseline* annotations from the full set of 3,984 sentences.

3.4. Evaluation data

In order for a meaningful comparison between the crowd and expert models, the evaluation set needs to be carefully vetted. For each of the relations, we started by selecting the positive/negative threshold for *sentence-relation score* such that the crowd agrees the most with the experts. We assume that, if both the expert and the crowd agree that a sentence is either a positive or negative example, it can automatically be used as part of the test set. Such a sentence was labeled with the expert score.

The interesting cases appear when crowd and expert disagree. To ensure a fair comparison, our team adjudicated each of them to decide whether or not the relation is present in the sentence. The sentences where no decision could be reached were subsequently removed from the evaluation. This set was constituted of confusing and ambiguous sentences that our team could not agree on, often because there was some lexical problem with the sentence (such as enumerations, term span problems, etc.). Eliminating these sentences is a disadvantage to a system like ours which was motivated specifically by the need to handle such cases, however the scientific community still only recognizes discrete measures such as precision and recall, and we felt it only fair to eliminate the cases where we could not agree on the correct way to map ambiguity into a discrete score.

We collected adjudicated test labels for all the 902 sentences used in the first experiment. For evaluation, we selected sentences through 5-fold cross-validation, but we obviously only used the test labels when a partition was chosen to be test. For the second evaluation over 3,984 sentences, we again selected test sets using cross-validation over the 902 sentences, adding the unselected sentences with their training labels to the training set. This allows us to directly compare the learning curves between the 902 and 3,984 experiments. The scores reported are the mean over the cross-validation runs.

3.5. CrowdTruth-weighted evaluation

We also explored how to incorporate CrowdTruth into the evaluation process. The reasoning of our approach is that the ambiguity of a sentence should also be accounted for in the evaluation – i.e. sentences that do not clearly express a relation should not count for as much as clear sentences. In this case, the *sentence-relation score* gives a real-valued score that measures the degree to which a particular sentence expresses a particular relation between two terms. Therefore, we propose a set of evaluation metrics that have been weighted with the *sentence-relation score* for a given relation. The

metrics have been previously tested on a subset of our ground truth data, as detailed in [Dumitrache et al. 2015a].

We collect true and false positives and negatives in the standard way, such that $tp(s) = 1$ iff s is a true positive, and 0 otherwise, similarly for fp, tn, fn . The positive sentences (i.e true positive and false negative labels) are weighted with the sentence-relation score $srs(s)$ for the given sentence-relation pair, i.e. the likelihood the relation is expressed in the sentence. Negative sentences (true negative and false positive labels) are weighted with $1 - srs(s)$, the likelihood that the sentence does not express the relation. Based on this, we define the following metrics to be used in the evaluation:

— *weighted precision*: Where normally $P = tp/(tp + fp)$, weighted precision

$$P' = \frac{\sum_s srs(s) \cdot tp(s)}{\sum_s srs(s) \cdot tp(s) + (1 - srs(s)) \cdot fp(s)};$$

— *weighted recall*: Where normally $R = tp/(tp + fn)$, weighted recall

$$R' = \frac{\sum_s srs(s) \cdot tp(s)}{\sum_s srs(s) \cdot tp(s) + srs(s) \cdot fn(s)};$$

— *weighted F-measure*: Is the harmonic mean of weighted precision and recall:

$$F1' = 2P'R'/(P' + R').$$

4. RESULTS

4.1. CrowdTruth vs. medical experts

In the first experiment, we compare the quality of the crowd with expert annotations over 902 sentences (those that have been also annotated by experts). We start by comparing the crowd and expert labels to the adjudicated test labels on each sentence, without training a classifier, computing an F1 score that measures the *annotation quality* of each set, shown in Fig.1 & 2. Since the baseline, expert, and single sets are binary decisions, they appear as horizontal lines, whereas the crowd annotations are shown at different sentence-relation score thresholds. For both relations, the crowd labels have the highest annotation quality F1 scores, 0.907 for the *cause* relation, and 0.966 for *treat*. The expert data is close behind, with an F1 score of 0.844 for *cause* and 0.912 for *treat*. McNemar's test [McNemar 1947] over paired nominal data was used to calculate statistical significance of the results. This difference between crowd and expert is significant with $p = 0.007$ for *cause*, and $p = 0.023$ for *treat*. The sentence – relation score threshold for the best annotation quality F1 is also the threshold where the highest agreement between crowd and expert occurs (Fig.3 & 4).

Next we compare the quality of the crowd and expert annotations by training the relation extraction model. For the *cause* relation, the results of the evaluation (Fig.5) show the best performance for the crowd model when the sentence-relation threshold is 0.5. Trained with this data, the classifier model achieves an F1 score of 0.642, compared to the expert-trained model which reaches 0.638. The difference is statistically significant with $p = 0.016$.

Tab.III shows the full results of the evaluation, together with the results of the CrowdTruth weighted metrics (P' , R' , $F1'$). In all cases, the $F1'$ score is greater than F1, indicating that ambiguous sentences have a strong impact on the performance of the classifier. Weighted P' and R' also have higher values in comparison with simple precision and recall.

For the *treat* relation, the results of the evaluation (Fig.6) shows baseline as having the best performance, at an F1 score of 0.856. The crowd dataset, with an F1 score of 0.854, still out-performs the expert, scoring at 0.832. These three scores are not,

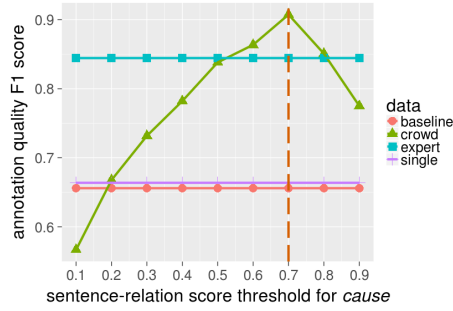


Fig. 1: Annotation quality F1 scores for the *cause* relation.

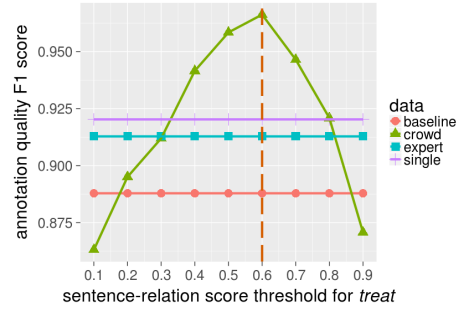


Fig. 2: Annotation quality F1 scores for the *treat* relation.

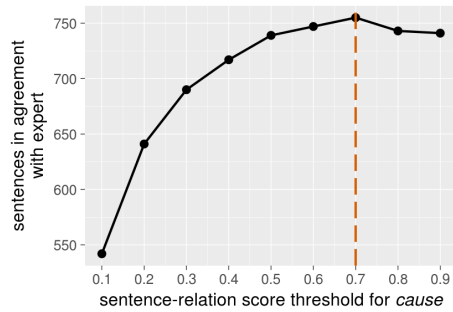


Fig. 3: Crowd & expert agreement for the *cause* relation.

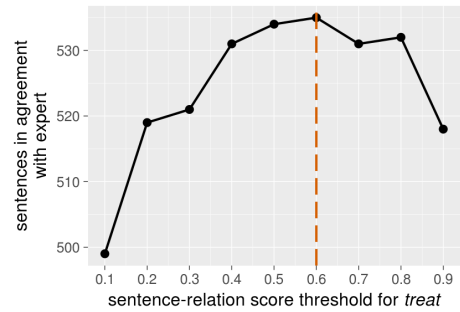


Fig. 4: Crowd & expert agreement for the *treat* relation.

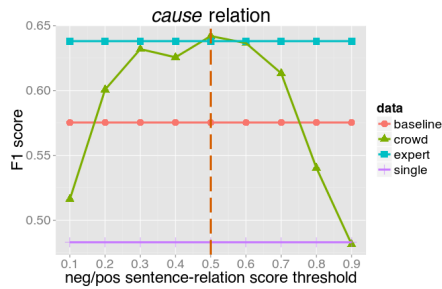


Fig. 5: Model testing F1 scores for the *cause* relation.

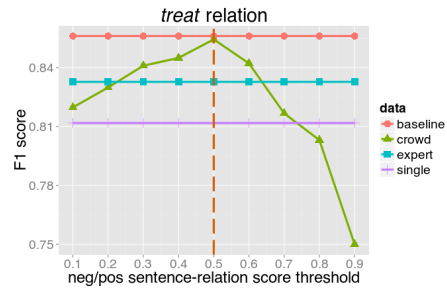


Fig. 6: Model testing F1 scores for the *treat* relation.

however, significantly different ($p > 0.5$), as there are so few actual pairwise differences (a consequence of the higher scores and the size of the dataset).

For both *cause* and *treat* relations, the single annotator dataset performed the worst. It is also worth noting that the sentence – relation score threshold for the best classifier performance (0.5 for both relations) is different from the threshold for best annotation quality, and highest agreement with expert (0.7 for *cause* and 0.6 for *treat*, Fig.1 & 2).

Finally, we checked whether the number of workers per task was sufficient to produce a stable sentence-relation score. We did this in two ways, first by measuring the

Table III: Model evaluation results over 902 sentences. Crowd scores are shown at 0.5 negative/positive sentence-relation score threshold.

	Dataset	P	P'	R	R'	F1	F1'
<i>cause</i> relation	<i>crowd</i>	0.565	0.632	0.743	0.754	0.642	0.687
	<i>expert</i>	0.672	0.711	0.604	0.616	0.638	0.658
	<i>baseline</i>	0.436	0.474	0.844	0.842	0.575	0.606
	<i>single</i>	0.495	0.545	0.473	0.478	0.483	0.658
<i>treat</i> relation	<i>crowd</i>	0.823	0.843	0.891	0.902	0.854	0.869
	<i>expert</i>	0.834	0.863	0.833	0.84	0.832	0.85
	<i>baseline</i>	0.767	0.811	0.968	0.968	0.856	0.882
	<i>single</i>	0.774	0.819	0.856	0.866	0.811	0.84

cosine distance between the sentence vectors at each incremental number of workers (Fig. 7), and second by measuring the annotation quality F1 score against the number of workers annotating each sentence (Fig. 8). The plot of the mean cosine distance between sentence vectors before and after adding the latest worker shows that the sentence vector becomes stable after 10 workers. The annotation quality F1 score per total number of workers is also stable after 10 workers (the drop towards the end is due to sparse data – only 54 sentences had 15 or more total workers). Note also that in both charts the red lines show the number of sentences that have been annotated by at least that number of workers, and are plotted against the right axis.

As a result of this analysis, we ensured that each sentence was annotated by at least 10 workers, after spam removal.

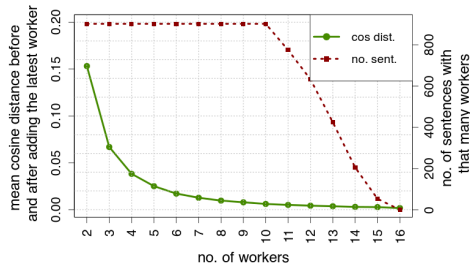


Fig. 7: Mean cosine distance for sentence vectors before and after adding the latest worker, shown per number of workers.

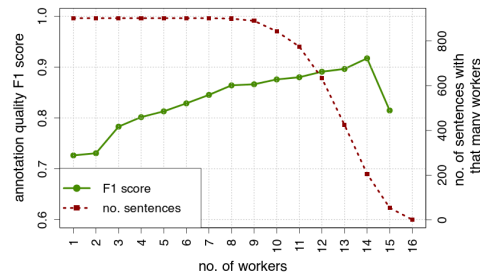


Fig. 8: Annotation quality F1 for crowd pos./neg. threshold at 0.7, shown per number of workers.

4.2. CrowdTruth vs. distant supervision

Distant supervision is a widely used technique in NLP, because its obvious flaws can be overcome at scale. We did not have enough time with the experts to gather a larger dataset from them, but the crowd is always available, so after we determined that the performance of the crowd matched the medical experts, we extended the experiments to 3,984 sentences. The crowd dataset in this experiment uses a fixed sentence-relation score threshold equal to 0.5, since this is the value where the crowd performed the best in the previous experiment, for both of the relations. As in the previous experiment, we employed five-fold cross validation to train the model. The test sets were kept the same as in the previous experiment, using the test partition labels as a gold standard. The

goal was to compare the crowd to the distant supervision baseline, while scaling the number of training examples, until achieving a stable learning curve in the F1 score. Since the single annotator dataset performed badly in the initial experiment, it was dropped from this analysis. The full results of the experiment are available in Tab.IV.

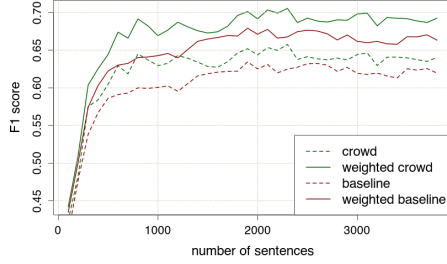


Fig. 9: Learning curves for *cause* relation.

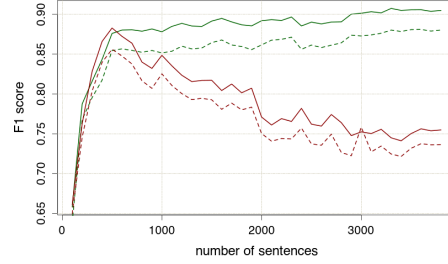


Fig. 10: Learning curves for *treat* relation.

Table IV: Model evaluation results over 3,984 sentences. Crowd scores are shown at 0.5 sentence-relation score threshold.

	Dataset	P	P'	R	R'	F1	F1'
<i>cause</i> relation	<i>crowd</i>	0.538	0.61	0.79	0.802	0.64	0.692
	<i>baseline</i>	0.475	0.53	0.889	0.887	0.619	0.663
<i>treat</i> relation	<i>crowd</i>	0.876	0.913	0.887	0.898	0.88	0.904
	<i>baseline</i>	0.808	0.858	0.678	0.673	0.736	0.754

For both relations, the crowd consistently performs better than the baseline. In the case of the *cause* relation, crowd and baseline perform closer to each other, with an F1 score of 0.64 for crowd and 0.619 for baseline. This difference is significant with $p = 0.001$. The gap in performance is even greater for accuracy, where the crowd model scored at 0.773 and baseline at 0.705. The learning curves for the *cause* relation (Fig.9) show both datasets achieve stable performance.

For the *treat* relation, the crowd scores an F1 of 0.88, while baseline scores 0.736, with $p = 1.39 \times 10^{-10}$ significance. The learning curves (Fig.10) show that, while baseline out-performed crowd when training with less than 1,000 sentences, crowd performance became stable after 1,000, while baseline went down, significantly increasing the gap between the two datasets.

The gap in performance is also present in the weighted F1' metrics. As is the case in the previous experiment, the F1' scores higher than the regular F1 score for both crowd and baseline. The only weighted metric that does not increase is the baseline recall. This is also the only metric by which the baseline model performed better for the regular F1 score.

5. DISCUSSION

5.1. CrowdTruth vs. medical experts

Our first goal was to demonstrate that, like the crowdsourced medical entity recognition work by [Zhai et al. 2013], the CrowdTruth approach of having multiple annotators with precise quality scores can be harnessed to create gold standard data with a quality that rivals annotated data created by medical experts. Our results show this clearly, in fact with slight improvements, with a sizable dataset (902 sentences) on a problem (relation extraction) that *prima facie* seems to require more domain expertise (than entity recognition).

The most interesting result of the first experiment is that the sentence-relation score threshold that gives the best F1 score is the same for both *cause* (Fig.5) and *treat* (Fig.6) relations, at a value of 0.5. This shows that ambiguous data is indeed valuable in training of clinical NLP models, and that being too strict with what constitutes a positive (or negative) training example produces flawed ground truth data. It is also worth noting that the single crowd annotator performs the worst for each of the relations. This could be further indication that the crowd can only achieve quality when accounting for the choices of multiple annotators, and further calls into question the standard practice of using only one annotator per example.

A curious aspect of the results is that the sentence-relation score threshold that gives the highest annotation quality F1 score (i.e. F1 score calculated directly over the test data, without training the model), shown in Fig.1 & 2, is different from the best threshold for classifier performance (Fig.5 & 6). It is the lower threshold (equal to 0.5) that results in the best model. This is most likely due to the higher recall of the lower threshold, which exposes the classifier to more positive examples. F-score is the harmonic mean between precision and recall, and does not necessarily represent the best trade-off between them, as this experiment shows for annotation quality. Indeed F-score may not be the best trade-off between precision and recall for the classifier, either, but it is the most widely accepted and reported metric for relation extraction. Note also that for both relations, the annotation quality at the 0.5 threshold is comparable or better than expert annotation quality.

In our error analysis of the annotation quality, we found that (as Figs. 1 & 2 show) experts and the crowd both make errors, but of different kinds. Experts tend to see relations that they know hold as being expressed in sentences, when they are not. For example, in, “He was the first to describe the relation between **Hemophelia** and **Hemophilic Arthropathy**,” experts labeled the sentence as expressing the *cause* relation, since they know Hemophelia causes Hemophilic Arthropathy. Thus they are particularly prone to errors in sentences selected by distant supervision, since that is the selection criterion. Crowd workers, on the other hand, were more easily fooled by sentences that expressed one of the target relations, but *not between the selected arguments*. For example, in “Influenza treatments such as **antivirals** and **antibiotics** are sometimes recommended,” some crowd workers will label the sentence with treats, even though we are looking for the relation between *antivirals* and *antibiotics*. The crowd achieves overall higher annotation quality due to redundancy, over the set of 10 workers, it is unlikely they will all make the same mistake.

In Figs. 7 & 8 we observe that we need at least 10 workers to get a stable crowd score. This result goes against the general practice for building a ground truth, where per task there usually are 1 to 5 annotators. Based on our results, we believe that the general practice is wrong, and that outside of a few clear cases, the input of more annotators per task is necessary to account for the ambiguities inherent in language, as well as all other interpretation tasks (e.g. images, audio, event processing, etc.). Even with this added requirement, we found that crowd data is still cheaper to ac-

quire than annotation from medical experts, as the crowd is both cheap (the cost of the experts was 50% higher) and always available via dedicated crowdsourcing platforms like CrowdFlower.

A bottleneck in this analysis is the availability of expert annotations – we did not have the resources to collect a larger expert dataset, and this indeed is the main reason to consider crowdsourcing. In this context, the real value of distant supervision is that large amounts of data can be gathered rather easily and cheaply, since humans are not involved. Therefore, the goal of the second experiment was to explore the trade-off between quality and cost of crowdsourcing compared to distant supervision, while scaling up the model to reach its maximum performance.

5.2. CrowdTruth vs. distant supervision

The results for both relations (Fig.9 & Fig.10) show that the crowd does out-perform the distant supervision baseline after the learning curves have stabilized, thus justifying its cost in terms of money. From this we infer that not only is the crowd generating higher quality data than the automated baseline, but training the model with weights, as opposed to binary labels, does have a positive impact on the performance of the model.

The results of the CrowdTruth weighted F1' consistently scored above the simple F1, for both baseline and crowd over both relations. This consolidates our assumption that ambiguity does have an impact on classifier performance, and weighting test data with ambiguity can account for this hidden variable in the evaluation.

The only weighted metric without a score increase is the baseline R' for the *cause* relation (see Tab.IV). Recall is also the only un-weighted metric for which the *cause* baseline model performed better than the crowd. Recall is inversely proportional to the number of false negatives, indicating that distant supervision, for this relation, is finding more positives at the expense of incorrectly labeling some of them. This appears to be a consequence of how the model performs its training – one of the features it learns is the UMLS type of the terms. For the *cause* relation, it seems that term types are often enough to accurately classify a positive example (e.g. an anatomical component will rarely be the effect of a causal relation).

Over-fitting on term types classification could also be the reason that baseline performs better than the crowd in the initial experiment for *treat* (Tab.III), where recall for baseline is unusually high. *treat* is also a relation that appears to favor a high recall approach – there are very few negative examples where the type constraint of the terms (drug - disease) is satisfied. In previous work [Aroyo and Welty 2014] we observed that *treat* generates less ambiguity than *cause*, which explains why *treat* has overall higher F1 scores than *cause* in all datasets. However, the high F1 scores could also make the models for *treat* more sensitive to confusion from ambiguous examples, as a small number of confusing sentences would be enough to decrease such a high performance. Indeed, as more (potentially ambiguous) examples appear in the training set, both the F1 and the recall of the baseline for *treat* drop, while the crowd scores remain consistent (Fig.10). This result emphasizes the importance of weighting training data with ambiguity, as a few ambiguous examples seem to have a strong impact in generating false negatives during classification.

6. CONCLUSION

The standard data labeling practice used in supervised machine learning attempts to minimize disagreement between annotators, and therefore fails to model the ambiguity inherent in language. We propose the CrowdTruth method for collecting ground truth through crowdsourcing, that reconsiders the role of people in machine learning

based on the observation that disagreement between annotators can signal ambiguity in the text.

In this work, we used CrowdTruth to build a gold standard of 3,984 sentences for medical relation extraction, focusing on the *cause* and *treat* relations, and used the crowd data to train a classification model. We have shown that, with the processing of ambiguity, the crowd performs just as well as medical experts in terms of the quality and efficacy of annotations, while being cheaper and more readily available. In addition, our results show that, when the model reaches maximum performance after training, the crowd also performs better than distant supervision. Finally, we introduced and validated new weighted measures for precision, recall, and F-measure, that account for ambiguity in both human and machine performance on this task. These results encourage us to continue our experiments by replicating this methodology for an increasing set of relations in the medical domain.

Acknowledgments

The authors would like to thank Dr. Chang Wang for support with using the medical relation extraction classifier, and Anthony Levas for help with collecting the expert annotations. The authors, Dr. Wang and Mr. Levas were all employees of IBM Research when the expert data collection was performed, and we are grateful to IBM for making the data freely available subsequently.

REFERENCES

- Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 17.
- Lora Aroyo and Chris Welty. 2013a. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *Web Science 2013*. ACM (2013).
- Lora Aroyo and Chris Welty. 2013b. Measuring crowd truth for medical relation extraction. In *AAAI 2013 Fall Symposium on Semantics for Big Data*.
- Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34. Issue 1. DOI: <http://dx.doi.org/10.15346/hc.v1i1.3>
- Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What Determines Inter-coder Agreement in Manual Annotations? A Meta-analytic Investigation. *Comput. Linguist.* 37, 4 (Dec. 2011), 699–725. DOI: <http://dx.doi.org/10.1162/COLI.a.00074>
- Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl 1 (2004), D267–D270.
- John D Burger, Emily Doughty, Sam Bayer, David Tresner-Kirsch, Ben Wellner, John Aberdeen, Kyungjoon Lee, Maricel G Kann, and Lynette Hirschman. 2012. Validating candidate gene-mutation relations in MEDLINE abstracts via crowdsourcing. In *Data Integration in the Life Sciences*. Springer, 83–91.
- Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D’Avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association* 18, 5 (2011), 540–543.
- Michelle Cheatham and Pascal Hitzler. 2014. Conference v2. 0: An uncertain version of the OAEI Conference benchmark. In *The Semantic Web—ISWC 2014*. Springer, 33–48.
- David L Chen and William B Dolan. 2011. Building a persistent workforce on mechanical turk for multilingual data collection. In *Proceedings of The 3rd Human Computation Workshop (HCOMP 2011)*.
- Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’13)*. ACM, New York, NY, USA, 1999–2008. DOI: <http://dx.doi.org/10.1145/2470654.2466265>
- Anca Dumitrache, Lora Aroyo, and Chris Welty. 2015a. Achieving Expert-Level Annotation Quality with CrowdTruth: the Case of Medical Relation Extraction. In *Proceedings of Biomedical Data Mining, Modeling, and Semantic Integration (BDM2I) Workshop, International Semantic Web Conference (ISWC) 2015*.

- Anca Dumitrache, Lora Aroyo, and Chris Welty. 2015b. CrowdTruth Measures for Language Ambiguity: the Case of Medical Relation Extraction. In *Proceedings of Linked data for Information Extraction (LD4IE) Workshop, International Semantic Web Conference (ISWC) 2015*.
- Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *In Proc. NAACL HLT (CSLDAMT '10)*. Association for Computational Linguistics, 80–88.
- Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 377–382.
- Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. In *The Semantic Web–ISWC 2014*. Springer, 486–504.
- Sarath Kumar Kondreddi, Peter Triantafillou, and Gerhard Weikum. 2014. Combining information extraction and human computing for crowdsourced knowledge acquisition. In *30th International Conference on Data Engineering*. IEEE, 988–999.
- Jey Han Lau, Alexander Clark, and Shalom Lappin. 2014. Measuring gradience in speakers grammaticality judgements. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. 821–826.
- Jongwuk Lee, Hyunsouk Cho, Jin-Woo Park, Young-rok Cha, Seung-won Hwang, Zaiqing Nie, and Ji-Rong Wen. 2013. Hybrid entity clustering using crowds and data. *The VLDB Journal* 22, 5 (2013), 711–726. DOI: <http://dx.doi.org/10.1007/s00778-013-0328-8>
- Tong Shu Li, Benjamin M Good, and Andrew I Su. 2015. Exposing ambiguities in a relation-extraction gold standard with crowdsourcing. *arXiv preprint arXiv:1505.06256* (2015).
- Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 2 (1947), 153–157.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*. Association for Computational Linguistics, 1003–1011.
- Jonathan M Mortensen, Mark A Musen, and Natalya F Noy. 2013. Crowdsourcing the verification of relationships in biomedical ontologies. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 1020.
- Charles Kay Ogden and I.A. Richards. 1923. *The meaning of meaning*. Trubner & Co, London.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong?. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 507–511.
- Mike Schaeckermann, Edith Law, Alex C. Williams, and William Callaghan. 2016. Resolvable vs. Irresolvable Ambiguity: A New Hybrid Framework for Dealing with Uncertain Ground Truth. In *1st Workshop on Human-Centered Machine Learning at SIGCHI 2016*.
- Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 254–263. <http://dl.acm.org/citation.cfm?id=1613715.1613751>
- Chang Wang and James Fan. 2014. Medical Relation Extraction with Manifold Models. In *52nd Annual Meeting of the ACL, vol. 1*. Association for Computational Linguistics, 828–838.
- Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. 2010. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*. 2424–2432.
- Chris Welty, James Fan, David Gondek, and Andrew Schlaikjer. 2010. Large Scale Relation Detection. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading (FAM-LbR '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 24–33. <http://dl.acm.org/citation.cfm?id=1866775.1866779>
- Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). Curran Associates, Inc., 2035–2043. <http://papers.nips.cc/paper/3644-whose-vote-should-count-more-optimal-integration-of-labels-from-labelers-of-unknown-expertise.pdf>

Haijun Zhai, Todd Lingren, Louise Deleger, Qi Li, Megan Kaiser, Laura Stoutenborough, and Imre Solti. 2013. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *JMIR* 15, 4 (2013).